We would like to thank the editor, Dr. Efrat Morin, and the two referees, Dr. Andreas Efstratiadis and an anonymous reviewer, for their constructive suggestions on our manuscript. The comments greatly improved our manuscript. Please find detailed answers to all the comments of the reviewers. We also enclose a word file in which all the changes are highlighted to be easily tracked.

Response to Dr. Efstratiadis

General comments:

- The paper is well-structured, well-written and easy to follow. I am very happy with the experience gained from this exhaustive modelling experiment, which reveals the superiority of pooled calibration (i.e. estimation of model parameters on the basis of flow data at multiple sites across the basin) over the stepwise strategy, and also reveals the advantages of semi-distributed over (non-parsimonious) fully-distributed schematizations, by means of improved predictive capacity and reduced parameter uncertainty. These outcomes are in agreement with the "holistic" approach proposed by Nalbantis et al. (2011) and other researchers of the same philosophy, which recognize that: (a) model complexity should be as high as allowed by the available information, and (b) all available information – even a single measurement – is valuable and should be accounted for in calibration. Unlikely, this is not the dominant philosophy among modellers, thus I believe that this paper will be a significant contribution to both hydrological science and practice. By reading this very good paper, I detected some issues to be clarified or further discussed, thus my recommendation is for a minor revision. In the following list, please find my specific comments as well as some technical corrections, to be addressed in your revision.

  We are very pleased that you enjoyed reading our manuscript and found it useful. We appreciate your encouraging and constructive comments. Below, please find our responses to all your specific comments and technical corrections.

Specific comments:

1. p. 10275, lines 15-17: "Importantly, distributed hydrologic models can evaluate hydrological response at interior ungaged sites, a benefit not afforded by conceptual, lumped models." Please, remove "conceptual", which refers to the modelling approach behind the formulation of the governing equations and not the spatial discretization of the model domain. Apart from lumped models, semi-distributed schemes are also by definition conceptual. Quoting Beven (1989), even a fully-distributed physically-based model can be regarded as conceptual, at the grid scale.

   We removed "conceptual" in that sentence.

2. p. 10275, line 27 to p. 10276, line 2: “Parameters can be discretized across the watershed in several ways: uniquely for each grid cell (fully distributed), based on hydrologic response units (semi-distributed), or in the simplest case, a single parameter set for all model grid cells (lumped).” In hydrologic models, hydrologic response units (HRUs) are mainly used for distributed and less often for semi-distributed schemes (e.g. Efstratiadis et al., 2008). The concept of HRUs was introduced by Flugel (1995) to characterize homogeneous areas with similar geomorphologic and hydrodynamic properties. The one-to-one correspondence of HRUs and sub-basins could be considered a specific case, which is however not consistent with the rationale of HRUs, as far as sub-basins have arbitrary boundaries that do not necessarily ensure homogenous characteristics.

   The sentence has been rewritten with new references as follows.
   “Parameters can be discretized across the watershed in several ways (Flugel, 1995; Efstratiadis et al., 2008; Khakbaz, et al., 2012): uniquely for each grid cell or hydrologic response unit (fully distributed), based on sub-basins whose boundaries do not necessarily ensure homogenous characteristics (semi-distributed), or in the simplest case, a single parameter set for all model grid cells (lumped).”

3. p. 10276, lines 26-30: “Many studies have reported that distributed models calibrated at the basin outlet are less accurate at interior locations (Anderson et al., 2001; Cao et al., 2006; Wang et al., 2012), but the extent of the error and uncertainty is unknown due to the computational expense needed to explore this issue.” To my opinion (and my experience), the accuracy of predictions of runoff at interior points mainly depends on the local characteristic of the basin. In the case of strongly heterogeneous basins, it is far from reasonable to make estimations based on the lumped information obtained at the basin outlet. On the other hand, if the key properties of the basin that influence runoff generation (e.g., permeability, vegetation, slope) do not vary significantly, such estimations could be quite reliable. However, the latter is not the rule.

   We agree with this point. The sentence has been rewritten as follows.
   “In the case of significant spatial variability in the basin properties that influence runoff generation (e.g., permeability, vegetation, slope, etc.), accurate runoff predictions are unlikely at interior locations based only on the lumped information obtained at the basin outlet (Anderson et al., 2001; Cao, et al., 2006; Breuer et al., 2009; Lerat et al., 2012; Simith et al., 2012; Wang, et al., 2012). The extent of this error and uncertainty is not well understood for heterogeneous basins due to the computational expense required to explore this issue.”

4. p. 10277, lines 1-2: “. . . for an alternative climate, which is required in climate change impact studies”. My impression is that climate change studies over broader areas refer to systematic deviations from the average climatic conditions, and not to “alternative climates”.

   Instead of using “alternative climate”, we used “possible future climate conditions”.
   Also, we changed the same term in the later part of the manuscript.

5. p. 10277, lines 22-24: "Water resources from the basin are shared by Afghanistan and Pakistan and serve as a water supply source for more than 20 million people." How significant are water abstractions in this basin? Are they accounted for in the modelling scheme? Are there any important regulations that modify the flow regime across the basin?

We completely agree with reviewer's concern about human interfere. The Kabul River has the largest flow of all of Afghanistan's rivers, but it can irrigate only a limited area because there is little land suitable for agriculture in the Afghan part of the basin (Ahmad and Wasiq, 2004) – for the most part, the river flows through mountainous or rocky areas. According to World Bank, (2010), about 2,927 km^2 (4.3% of the total basin area) is agricultural land and the average annual flow of the Kabul River is approximately 24,000 million cubic meters (MCM). Irrigation is a large water demand since the annual water demand estimate for the agricultural use is about 2,000 MCM, or about 8.3% of the total annual flow. In our hydrologic modelling process, the water consumed by irrigated croplands is implicitly accounted for by the evapotranspiration module. We note that the degree of irrigation impact during the time frame used for calibration (1960-1981) is likely much smaller than the current level.

The Naglu dam, which is located in the western part the Kabul River basin (upstream of the Daronta streamflow gage), forms the largest and most important storage among dams in the basin (World Bank, 2010). The live storage of the Naglu dam is 379 MCM. We expect that using monthly data for calibration somewhat reduces the bias from human interference, particularly the daily operations of Naglu dam. Nevertheless, the calibration results for the gage below this dam (Daronta), and to a lesser extent the basin outlet (Dakah), should be approached with caution. Given that a majority of the gages examined in this study are on an underdeveloped branch of the Kabul River, issues of human interference on calibration are somewhat mitigated. We also note that the poor performance at Daronta is likely due in part to the impacts of water abstraction and the operation of Naglu dam.

This information has been provided accordingly in the text.
"Similar to most other hydrological models (Efstratisdis et al., 2008), HYMOD_DS is not designed to model water abstractions for agricultural lands and dam operations within the basin. According to World Bank (2010), water demand for agricultural use is about 2,000 MCM (million cubic meters), or about 8.3% of the total annual flow. The Naglu dam (Figure 1) upstream of the Daronta streamflow gage forms the largest and most important reservoir in the basin, with an active storage of 379 MCM. In our hydrologic modelling process, the water consumed by irrigated croplands is implicitly accounted for by the evapotranspiration module. We note that the degree of irrigation impact during the time frame used for calibration (1960-1981) is likely much smaller than the current level. We also expect that using monthly data for calibration somewhat reduces the bias from human interference, particularly the daily operations of Naglu dam. Nevertheless, the calibration results for the gage below this dam (Daronta), and to a lesser extent the basin outlet (Dakah), should be approached with caution. Given that a majority of the gages examined in this study are on an underdeveloped branch of the Kabul River, issues of human interference on calibration are somewhat mitigated."

6. Section 2 (Study area): Here you should add information about the flow stations and the available data, and also provide synoptic statistical information about the hydrological characteristics of the basin, e.g. mean annual flow at the seven stations of interest, mean precipitation over the sub-basins, etc. (you can add these data to Table 1). It would also be useful to refer to the physiographic properties of the basin and the dominant runoff mechanisms, which are essential to interpret the model results and plausibility of the optimized parameter values. It is also essential to explain to which extend is this basin heterogeneous, thus justifying the implementation of each parameterization approach and better explain the model results.

   Table 1 has been updated with basin climate information (mean annual precipitation, mean temperature, and flow) and geographic properties (drainage area, glacier area, mean elevation). This additional information is also discussed in the section describing the study area.

   "The streamflow regime of the Kabul River can be classified as glacial with maximum streamflow in June or July and minimum streamflow during the winter season. Approximately 70% of annual precipitation (475mm) falls during the winter season (November to April). While the dominant source of streamflow in winter is baseflow and winter rainfall, glaciers and snow cover are the most important long-term forms of water storage and, hence, the main source of runoff during the ablation period for the basin (Shakir et al., 2010). In total 2.9% (1954km^2) of the basin is glacierized based on the Randolph Glacier Inventory version 3.2 (Pfeffer, et al., 2014). The melt water from glaciers and snow produce the majority (75%) of the total streamflow (Hewitt, et al., 1989). Table 1 provides the climates and geophysical properties of each sub-watershed delineated by the stations located inside the Kabul Basin (Figure 1). Two different climate patterns are distinguishable across the sub-basins. The sub-basins on the Kunar River tributary (Kama, Asmar, Chitral, Gawardesh, and Chaghasarai) receive moderate annual precipitation and are highly affected by snow and glacier covers. All of these sub-basins have high ratios of mean annual flow to mean annual precipitation, with the ratios for the Kama, Asmar, Chitral, and Chaghasarai sub-basins larger than 1. Conversely, the Daronta sub-basin contains only minimal glacial cover, and is relatively dry. Daronta is also much less productive, with annual streamflow far below the other sub-basins with an average of only 165 mm/year."

7. p. 10279, lines 4-6: "However, in this particular study daily hydrologic model simulations can only be compared against available monthly streamflow records". It is not clear whether monthly streamflows are averaged values of daily (or hourly) observations or instantaneous values, gather e.g. from direct flow measurements. Such clarification is very important.

   Unfortunately, the only observations that are available for public use are monthly. There is a report (Olson and Williams-Sether, 2010) clarifying that each monthly streamflow is the mean of the daily values for the month, and monthly values are calculated from daily values for all complete months of record. However, the daily values are not made available because there are political issues surrounding the trans-boundary use of the river's waters and potential projects planned on the river.

We have added the following details in the manuscript to clarify the immediate question regarding the data.
"Streamflow data were not collected in Afghanistan after September 1980 until recently because streamgaging was discontinued soon after the Soviet invasion of Afghanistan in 1979 (Olson and Williams-Sether, 2010). Though measurements were taken at a daily time step, data are only made available for public use at monthly aggregated levels, calculated using the mean of the daily values."

8. p. 10279, lines 18-20: "No matter the parameterization scheme, the model structure follows the climate input grids, i.e. the hydrological water cycle within each grid cell is modelled separately." In the revised paper, I suggest also employing the simplest of model configurations, assuming a lumped structure for both model inputs and parameters (i.e. using the averaged precipitation over the basin). This classical lumped approach considering 15 (or less) parameters would provide, in theory, the optimal results at the basin outlet with minimal computational burden, to be considered as "baseline scenario".

We understand the reviewer's suggestion, and initially considered this ourselves. However, we wanted the comparisons in this paper to isolate the effects of calibration uncertainty rather than address the structural uncertainties surrounding the model grid distribution (or lack thereof). Also, since a large focus of our study is on ungaged, interior point streamflow estimation, a lumped model structure would not really be appropriate (unless there was some scaling from flow estimates at the outlet of the basin to the interior points). With that said, we do agree with the reviewer that this issue should at least be addressed. Therefore, we now include a preliminary test of the basin outlet model from the lumped HYMOD without the gridded structure (13 parameters and basin-averaged climate inputs). The 2 parameters associated with the river routing models are dropped due to its lumped structure. We have added a description regarding a preliminary performance comparison between this model and its analogue with a gridded structure. Since the distributed model outperformed, we used this as a justification to set our "baseline" model as having a distributed structure. We decided that a figure is necessary for this additional part and a new figure has been provided as another supplementary material (Figure S3). We summarize these details in a new paragraph in the manuscript.

New paragraph:
"We note that a lumped model structure (i.e., no gridded or sub-unit structure) has often been considered as a baseline model formulation in the assessment of distributed modelling frameworks (e.g., see Simith et al., 2013). However, the focus of our study is on ungaged interior site streamflow estimation, making this formation somewhat inappropriate. Further, preliminary tests comparing streamflow simulations at the basin outlet (Dakah) between a gridded and basin-averaged structure, both with a lumped parameter formation, support the use of the distributed grid structure (Figure S3)"

New figure:

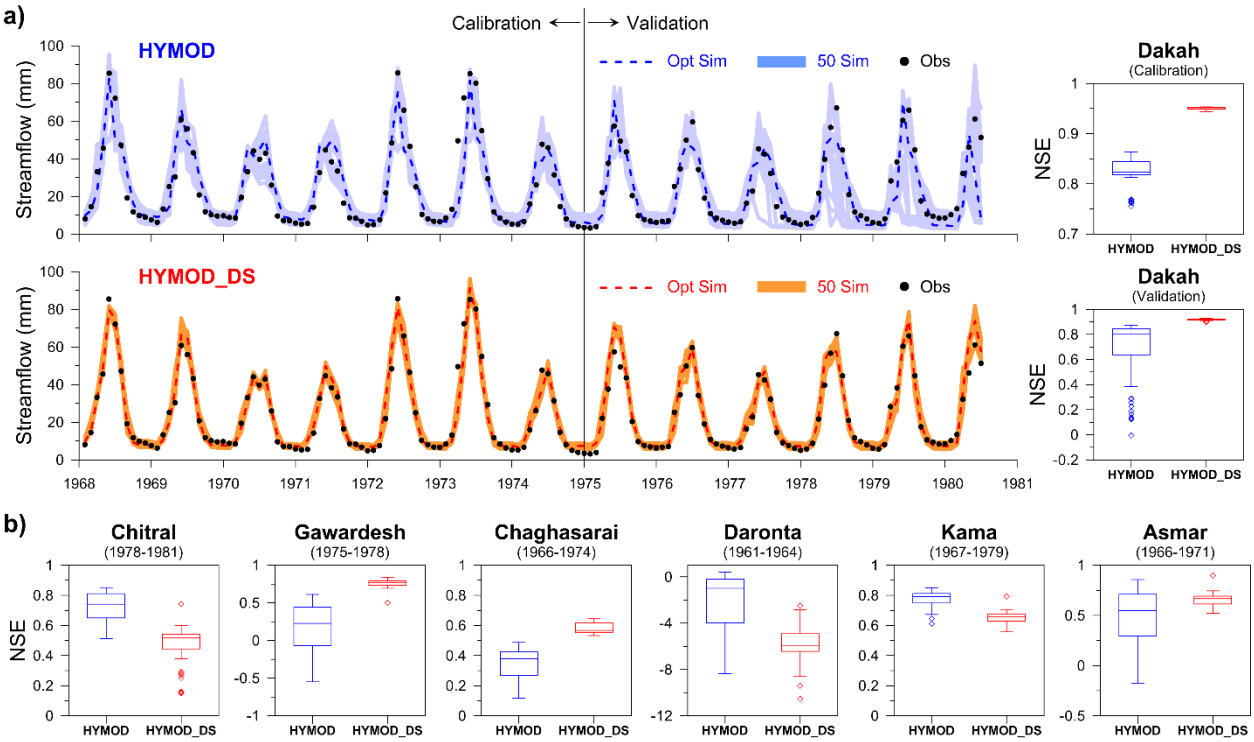


Figure S3. (a) Basin outlet (Dakah) simulations of HYMOD and MYMOD_DS (with the lumped parameterization) from 50 trials of calibration. The Box plots provide the performance evaluation on 50 simulations of both models for both calibration and validation periods. (b) Performances of the models at the interior points of the watershed are assessed.

9. p. 10279, lines 21-24: "The parameter complexity will vary depending on the calibration experiment being conducted, but for each experiment regardless of the parameterization, the optimization is implemented 50 times using the GA algorithm to explore parameter uncertainty." Parameter uncertainty is a combined effect of multiple causes, one of which is inefficient calibrations (i.e. calibrations trapped to local optima). Even the use of robust and sophisticated evolutionary algorithms cannot remedy this problem, especially when a large number of parameters are considered. However, there are also other sources of parameter uncertainty, associated with data errors, unknown boundary conditions, etc. In this context, I propose avoiding the general term "parameter uncertainty" and focus to "calibration uncertainty", which is very well represented in your work, by implementing 50 independent runs for each optimization problem.

   Throughout the manuscript, we replaced the term "parameter uncertainty" with "calibration uncertainty".

10. Section 3.1 (Multisite calibration): There are some important issues that are mentioned in next parts of the document, yet they should be also highlighted in this section. In order to better follow the modelling experiment, is essential to explain the sequence of sub-

catchments, which strongly affects the outcomes of stepwise calibration (thus I propose moving Fig. S1 from the supplement to the main text). Another missing issue is the lack of overlapping data periods among most of stations, which is a bad coincidence, since this weakens the multisite calibration approach: in fact, you do not have simultaneous information on the basin responses, which would allow account for the heterogeneity of the associated hydrological processes.

First, we have changed the paper structure to address some of these issues. Now, the Section "Data and Models" is placed ahead of Section "Methods" so that the reader sees some pertinent information regarding the basin before being introduced to the details of the modeling experiments. In the methods section, we moved Fig. S1 (now Figure 5) to the main text in section 4.1 (multisite calibration) to make sure the reader understands the sequence of sub-basins in the stepwise calibration. Also, at the end of this section, we now include a brief discussion of the second point made by the reviewer:

"It is important to note that the evaluation of these multisite calibration strategies is somewhat weakened because of the lack of overlapping data periods among most of the stations (Figure 2). This drawback prevents the calibration methods from accounting for simultaneous information from different tributaries, which, if available, would better enable the calibration methods to account for heterogeneity of hydrological processes across the sub-basins."

11. p. 10281, lines 23-27: ".. the lumped version of the HYMOD_DS contains a single, 15-member parameter set applied to all model grid cells. The semi-distributed conceptualization of HYMOD_DS contains a single parameter set for each sub-basin, totaling 75 parameters. In the distributed parameterization ... the number of parameters requiring calibration reaches 2400." Here it is worthy reminding that for the transformation of rainfall to hydrograph at the basin outlet, only 5 to 6 parameters can be identified on the basis of a single observation set (cf. Wagener et al., 2001). Under this premise, the number of parameters for the lumped scheme is realistic, taking into account that snow, glacier and flow routing processes are also modelled. For the semi-distributed approach, the number of parameters remains realistic, since external information is increased by accounting for interior flow data in calibrations. However, the distributed approach, with 2400 parameters to be optimized, is far from acceptable, and any attempt to interpret the outcomes of calibration is unreasonable.

Thank you for pointing out this issue and the useful references. We expanded our discussion section with this issue.

"It is worth noting that for the transformation of rainfall to runoff, up to five or six parameters can be identified on the basis of a single hydrograph (Wagner et al., 2001). Under this premise, the number of the HYMOD_DS parameters being calibrated in the semi-distributed approach remains realistic, but the fully distributed parameterization scheme likely causes poor identifiability of the parameters. Thus, pursuing a parsimonious configuration (e.g. optimization for a small portion of the parameters) with an effort to increase the amount of

information (e.g. multivariable/multisite) is critical in the calibration of watershed system models (Gupta et al., 1998; Efstratiadis et al., 2008)."

12. p. 10284, lines 11-12: "Monthly streamflow observations for seven locations in the Kabul River basin (Fig. 1) were gathered between calendar years 1961–1980". The same equation with comment 6: why monthly flow data and how are these data extracted?

    Please refer to the answer for the comment 7.

13. p. 10285, line 14-15: "The overall model structure of the HYMOD_DS and its 15 parameters are described in Fig. 4 and Table 2 respectively." The feasible ranges that are employed for the model parameters are extremely large thus resulting to huge parameter uncertainty (at least, a priori uncertainty). For instance, the maximum soil moisture capacity ranges from 5 to 1500 mm. I would expect that an experienced hydrologist would propose much more narrow bounds, taking into account the physical interpretation of those parameters and the local characteristics of the specific study area. I strongly believe that a hydrological model is not a mathematical game, and calibration is not a black-box exercise. In contrast, model parameters should always have some correspondence to the physical properties of the basin, which is yet not reflected in this work. In addition, a substantial reduction of feasible ranges would be beneficial for the calibration effort, which is tremendous (1000 parallel processors running for 7 days!).

    Our main focus is to explore a variety of calibration strategies which becomes a computationally exhaustive task but can be implemented with the aid of parallel computing power. We noticed that there might be an advantage of having wide feasible parameter ranges; we can expect to avoid priori errors that could be caused by inappropriately narrowing down the ranges. We decided to embrace the computational cost owing to the wide parameter ranges and then try to solve this issue with the high computation power available from the MGHPCC.

    Nonetheless, this is a very good point which is worth a further discussion.
    "We also note the important role of experienced hydrologists in designing a parsimonious hydrologic calibration (e.g. Boyle et al., 2000). In this study, the feasible ranges of the HYMOD_DS parameters were kept wide (as is often done in automatic hydrologic calibrations) without consideration of the physical properties of the basin; the judgment of local hydrologic experts could help reduce the feasible ranges used during the calibration and thus contribute to a reduction of calibration uncertainty."

14. p. 10286, line 15: The Hamon method for PET estimations is not widely known. Please, provide one or two sentences with a very synoptic description of this method (rationale, input data). Is this method suitable for the climatic regime of the study area?

    We provided more information on the Hamon method with an additional equation. Please refer to the following for the changes made in the text:

"The potential evapotranspiration (PET) is derived based on the Hamon method (Hamon, 1961), in which daily PET in mm is computed as a function of daily mean temperature and hours of daylight:

$$\mathrm{PET} = Coeff \cdot 29.8 \cdot L_d \cdot \frac{0.611 \times \exp\left(17.27 \cdot T/(T+273.3)\right)}{T+273.3}$$

where, L_d is the daylight hours per day, T is the daily mean air temperature (^oC), and Coeff is a bias correction factor. The hours of daylight is calculated as a function of latitude and day of year based on the daylight length estimation model (CBM model) suggested by Forsythe et al. (1995)."

Is this method suitable for the climatic regime of the study area?

As explained, the Hamon is a temperature based method. Despite its simplicity relative to more input-detailed models, some studies identified the model as a method that produce satisfactory estimates of PET. Here's some examples. Vorosmarty et al. (1998) compared to 11 different PET models for a wide range of climatic conditions across the conterminous US and found that the Hamon model is comparable to more input-detailed models, such as the Shuttleworth-Wallance. In a study of 5 PET models for use with global water balance models Federer et al. (1996) found that estimates of PET from the Hamon model agreed with estimates from other models across a wide range of climates. From a comparison of six PET models, Lu et al. (2005) recommended the Hamon method for regional applications in the southeastern US based on the criteria of availability of input data and correlations with actual ET values.

15. p. 10290, lines 16-17: "High accuracy holds even under the Lump_Outlet, which is somewhat surprising given the spatial heterogeneity of the basin." I do not agree that this is a surprising conclusion. The lumped configuration of HYMOD_DS has 15 parameters, which are far from sufficient to represent hydrographs of any complexity.

We understand the point here. We have changed the wording accordingly to now read:

"High accuracy holds even under the Lump_Outlet, despite the spatial heterogeneity of the basin."

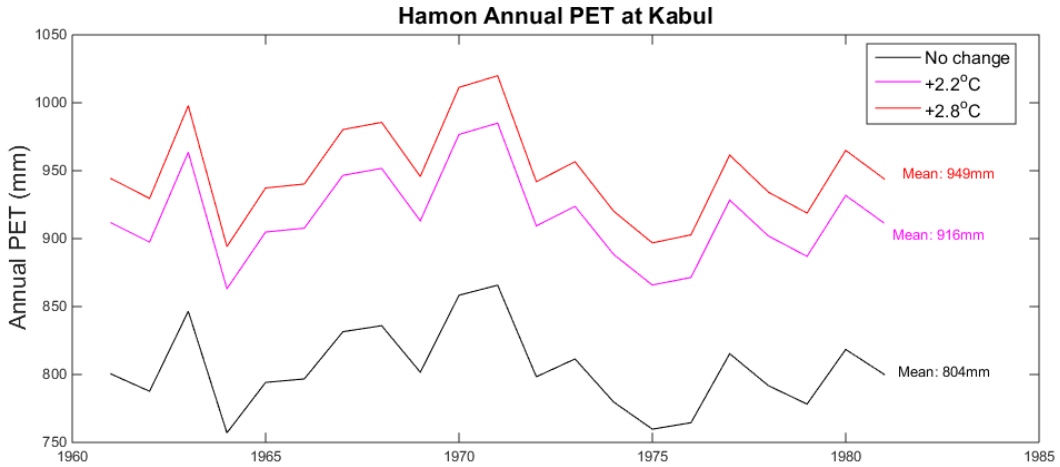
16. p. 10290, lines 25-27: ". . .the HYMOD_DS significantly overestimated streamflow at Daronta and underestimated flow at three sites in the eastern part of the basin" This is a strong evidence of the heterogeneity of the basin. Please, provide some information on the properties of the basin (e.g. geology) that would justify these differences.

We have updated Table 1 with the information to support the heterogeneity of the basin and also include new information on the basin heterogeneity in the updated Figure 1.
Please refer to the answer for the comment 6.

17. p. 10292, lines 7-9: “On the other hand, temperature clearly shows an upward trend for both radiative forcing scenarios. The average changes in annual temperature are +2.2oC and +2.8oC for RCP4.5 and RCP8.5, respectively”. Which are the impacts of such difference in PET estimations?

Changes in temperature are important in the PET estimation. The Hamon PET calculation is a function of temperature and daylight length. Since the daylight hours is a time-invariant variable, temperature changes will be the only factor affecting PET changes under the warming conditions. For an example, we took the grid cell covering Kabul city to calculate Hamon PET values under historical condition, +2.2^{o}C, and +2.8^{o}C. The value of calibrated Coeff (bias correction parameter) is 1.007 for this grid and the result is shown in the figure attached below. The average annual PET calculations are 804mm, 916mm, and 940mm under historical condition, +2.2^{o}C, and +2.8^{o}C, respectively. The percent changes of PET relative to the historical PET in the warming conditions of +2.2^{o}C and +2.8^{o}C are approximately +14% and +18%, respectively.

We have included a brief addition to the line in question for clarification:
“On the other hand, temperature clearly shows an upward trend for both radiative forcing scenarios. The average changes in annual temperature are +2.2^{o}C and +2.8^{o}C for RCP4.5 and RCP8.5, which, using the Hamon method, correspond to an increase in annual PET by approximately 100mm and 150mm, respectively.”



18. p. 10292, lines 17-19: “For the historical time period, all calibration schemes match the observed climatology at Dakah well, but monthly streamflow is underestimated in most of months at Kama and Asmar under the basin outlet calibrations”. If I understood well, you used as meteorological inputs the average projections of the 36 climate models during the period of observations. In that case, it is not clear whether the underestimation of monthly flows is due to inappropriate representation of past precipitation and temperature data by

climate models or due to inappropriate calibrations at the specific flow stations. For this reason, it is essential providing results on model bias (apart from NSE and KGE).

The 50 runs for the historical period have nothing to do with the climate model outputs. The observed climate is the only input that is used to derive the monthly streamflow estimates for the historical period with the 50 calibrated parameter sets. On the other hand, for the future period, 36 runs are related to a single parameter set because of 36 different GCM climate inputs. For each of the 50 parameter sets, we average out the uncertainty from the 36 future climate time series. In this way, the uncertainty ranges shown in Figure 10 both are composed of 50 different values, as described in the text "The whisker bars indicate the range across the 50 calibration trials; for the future scenarios, the whisker bars are derived by averaging over the 36 different climate projections for each of the 50 trials."

We have rewritten the part for a better clarification on this as follows.
"Figure 10 shows the monthly streamflow estimates for the historical period with the whisker bars indicating the uncertainty range across the 50 calibration trials. The monthly streamflow predictions are also provided for the 2050s under the RCP 4.5 and 8.5 scenarios. For the future scenarios, the whisker bars are derived by averaging over the 36 different climate projections for each of the 50 trials."

19. p. 10293, lines 26-27: "Another clear point is that the uncertainty resulting from different climate change scenarios substantially outweighs that from parameter uncertainty." This is of course a very important conclusion, and would deserve further discussion about the misuse of such scenarios as "deterministic" projections.

We discussed more about it in the section Discussion and Conclusion.
"We evaluated the separate and joint influence of uncertainties in parameter estimation and future climate on projections of seasonal streamflow and 100-year daily flood across calibration schemes and found that the uncertainty resulting from variations in projected climate between the CMIP5 GCMs substantially outweighs the calibration uncertainty. These results agree with other studies showing the dominance of GCM uncertainty in future hydrologic projections (Chen et al., 2011; Exbrayat et al., 2014). While the GCM-based simulations still have widespread use in assessing the impacts of climate change on water resources availability, the bounds of uncertainty resulting from an ensemble of GCMs cannot be well-defined because of the low credibility with which GCMs are able to produce timeseries of future climate (Koutsoyiannis et al., 2008). This issue hinders a straightforward appraisal of future water availability under climate change and has motivated other efforts; e.g. performance-based selection of GCMs (Perez et al., 2014)."

20. p. 10294, lines 10-14: "While no observed data is available against which to compare the results, an inter-model comparison is useful to distinguish the differences between the parameterization schemes." Since observed flood data are missing, these comparisons are little safe. You may use them in the context of a theoretical calibration exercise, but definitely not for decision-making purposes.

Yes, we agree. We changed the sentence as follows.
"Although the inter-model comparison is intended to be a useful addition that provides a distinction between the parameterization schemes in the pooled calibration approach, results from this analysis should be viewed in the context of a theoretical calibration exercise, not for decision-making purposes, because no observed daily streamflow is available against which to compare the estimated 100-year daily flood events."

Technical corrections:

1. p. 10278, lines 23, 24: Please, change to read "Sutcliffe".

   Done.

2. p. 10292, line 18: Term "observed climatology" is unclear. Climatology is defined as "the study of climate", while climate is defined as "as weather conditions averaged over a period of time" (http://en.wikipedia.org/wiki/Climatology).

   We changed it to "average monthly streamflow estimates"

3. p. 10292, line 21: Similarly, term "historical streamflow climatology" is not valid. I suppose that you refer to average monthly flow data?

   We changes it to "historical average monthly flow estimates"
   Throughout the manuscript we tried to correct the parts where the term "climatology" is used.

4. p. 10304, Table 1: Please, use common symbols for dates, e.g. YYYY/M or M/YYYY (not YYYY.M).

   Now it is in "YYYY/M"

5. p. 10316, Fig. 10: The coefficient of variation of which quantity is represented in the graphs? (similar for Fig. 12).

   For Fig. 10, it is for "Coefficient of variation of average season flow predictions"
   For Fig. 12, it is for "Coefficient of variation of 100-year flood estimations"
   We changed the y-axis label to reflect these clarifications in Fig. 10 and 12.
   Also, the captions for those figures are changed for more clear description of the figures.

Thank you.

Response to Anonymous Referee #2

General comments:

- I see one major limitation of the paper that leads me to ask for at least minor, if not major revisions: there is not much of a scientific discussion. The authors discuss their results most of all "with themselves" by comparing the various results they obtained. The discussion is short of any discussion with findings by other authors (e.g. on P10294 L3 the authors cite other work for the first time in the results and discussion section. This is on the last page of an eight pages long results and discussion section). There is plenty of published work about the effect of parameterization and their spatial variation, lumped vs distributed calibration approaches, performances of models in simulating interior gauges not considered in calibration, see for example results of the DMIP and LUCHEM projects, amongst others. Additionally, climate change effects on discharge in Central Asian catchments has been in the focus of many, many studies – how do these related to the results obtained here?

  Thank you for pointing out this. We also realized that there were not much discussion in the section "Results and Discussion". To try to follow the reviewer's suggestion, we expanded our discussion. First, we decided to focus on our results in the result section and change the paper's structure accordingly. Now we combined the discussion section with the conclusion part. Also, we expanded our discussion by introducing additional references in relevance to our work as suggested by the reviewer. Please find the revisions made in the section "Discussion and Conclusion" and detailed answers to all the specific comments in the following.

Specific comments:

- Title: High performance computing is mentioned in the title, but hardly presented in the method section, and not at all in the discussion. HPC in this paper is used as a technique to be able to run a large number of models, but it is not in the center of research as indicated by the title. I suggest to change the title.

  We understand your concern. We have changed the title to highlight our focus on a poorly gaged basin (which we feel is the more important emphasis of this work anyway). However, we do feel that the use of high performance computing is an important component of this work, so we tried to emphasize the necessity of exploiting parallel computing power to implement this kind of study in the abstract:

  "To address the research questions, high performance computing is utilized to manage the computational burden that results from high-dimensional optimization problems."

- P10276 L26 There are a number of papers which looked at model performance when excluding/including interior gauging stations during model calibration and validation; see e.g. the DMIP projects (Reed et al., 2004; Smith et al., 2012), the LUCHEM project (Breuer et al., 2009) or work by others (Andersen et al., 2001; Lerat et al., 2012).

  Thank you. We have added the recommended references.

- P10277 L1 You might want to have a closer look to a recent paper by Exbrayat et al. (2014) who investigated the contribution of uncertain model structures versus the impact of uncertain climate change projection to the global predictive model uncertainty. Even though not directly comparable to what the authors show here, it is worth considering and can be used in the discussion, which is lacking other researchers work (see general comment).

  Thank you for suggesting this useful reference. We expanded our discussion with the suggested reference.
  "These results agree with other studies showing the dominance of GCM uncertainty in future hydrologic projections (Chen et al., 2011; Exbrayat et al., 2014). …
  In addition to the uncertainties surrounding model parameters and future climate explored in this study, there is also significant uncertainty in streamflow projections stemming from structural differences between applied hydrologic models, which can be especially pertinent where robust calibration is hampered by the scarcity of data (Exbrayat et al., 2014). Further, the residual error variance of hydrologic model simulations would increase the effects of hydrologic model uncertainty as compared to that of the climate projections (Steinschneider et al., 2014). These issues need to be addressed in future work for exploring a comprehensive uncertainty assessment of climate change risk for poorly monitored hydrologic systems."

- P10277 L18 I do not agree that HPC is so new in hydrological modeling. I rather think that many researcher use HPC without highlighting it. Also in the work presented here, HPC is a tool that is used, but not a method that is further developed or presented in detail.

  We understand and have removed the language suggesting HPC is new in hydrological modeling. While we still feel that the use of HPC is uncommon and adds new possibilities for research questions, we agree that we are using HPC as a tool – it is not the focus of our study.

- P10278 L3 Is the annual precipitation 475 mm or are the 475 mm the 70% of total precipitation? Overall, the study area description is very short. Some more information about topography, soils/geology, flow characteristics, specific discharges from the subcatchments, and land use/management would be helpful to better understand some of the results.

  We dropped the number in the text to avoid any confusion caused by that. The number was meant to be for annul precipitation and is now provided in the updated Table 1.
  Figure 1 has been updated with more information (topography, soil types, and vegetation cover). We expanded the study area description accordingly.

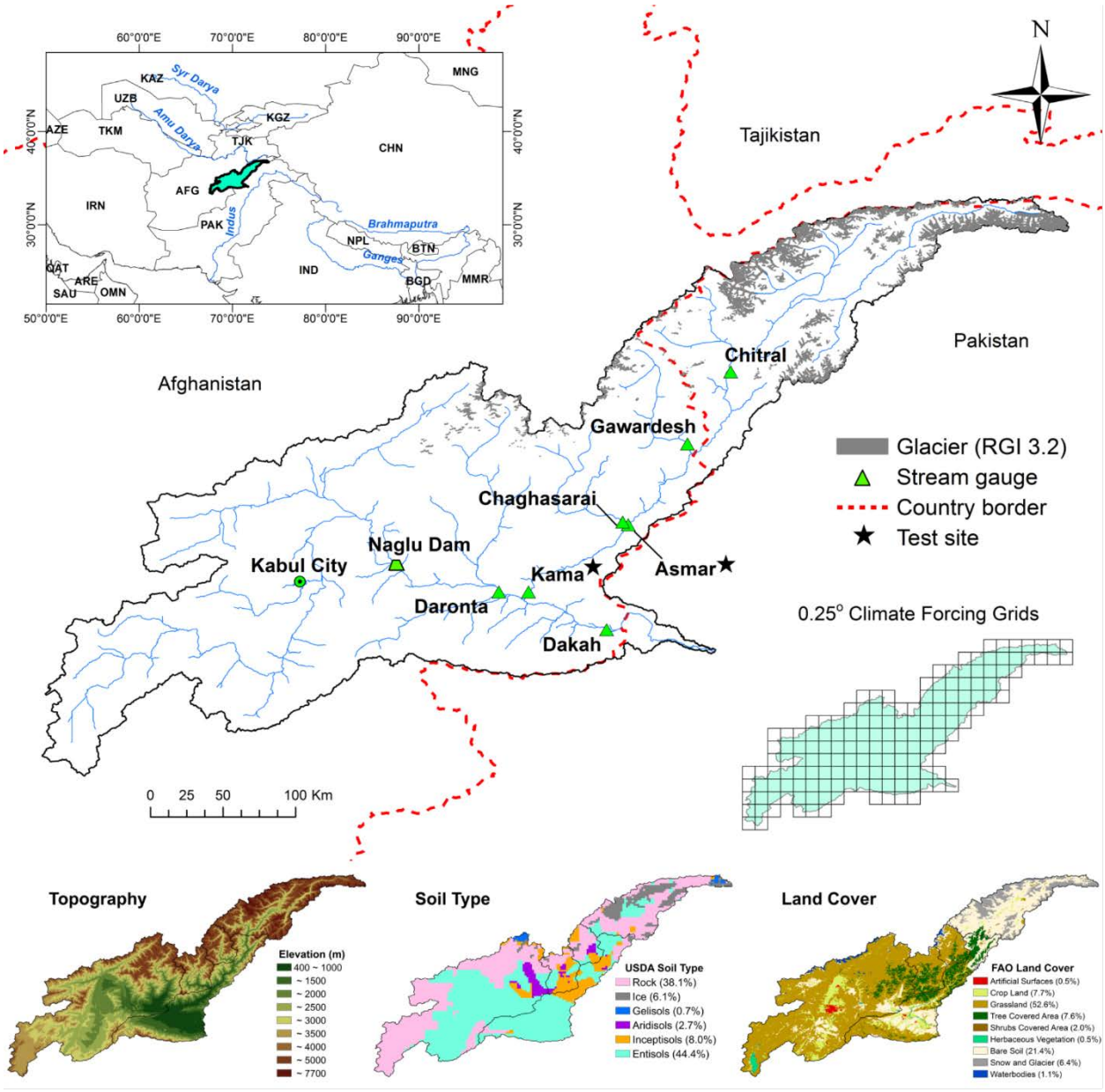


Figure 1. Kabul River Basin.

How about irrigation? Is it an important land management and if so, how did you deal with water abstraction. Looking at the often poor model performance in the western part of your catchment around Kabul I assume that missing information on water abstraction substantially influences your model performance.

We completely agree with reviewer's concern about human interfere. The Kabul River has the largest flow of all of Afghanistan's rivers, but it can irrigate only a limited area because there is little land suitable for agriculture in the Afghan part of the basin (Ahmad and Wasiq, 2004) – for the most part, the river flows through mountainous or rocky areas. According to World Bank, (2010), about 2,927 km^2 (4.3% of the total basin area) is agricultural land and the average

annual flow of the Kabul River is approximately 24,000 million cubic meters (MCM). Irrigation is a large water demand since the annual water demand estimate for the agricultural use is about 2,000 MCM, or about 8.3% of the total annual flow. In our hydrologic modelling process, the water consumed by irrigated croplands is implicitly accounted for by the evapotranspiration module. We note that the degree of irrigation impact during the time frame used for calibration (1960-1981) is likely much smaller than the current level.

The Naglu dam, which is located in the western part the Kabul River basin (upstream of the Daronta streamflow gage), forms the largest and most important storage among dams in the basin (World Bank, 2010). The live storage of the Naglu dam is 379 MCM. We expect that using monthly data for calibration somewhat reduces the bias from human interference, particularly the daily operations of Naglu dam. Nevertheless, the calibration results for the gage below this dam (Daronta), and to a lesser extent the basin outlet (Dakah), should be approached with caution. Given that a majority of the gages examined in this study are on an underdeveloped branch of the Kabul River, issues of human interference on calibration are somewhat mitigated. We also note that the poor performance at Daronta is likely due in part to the impacts of water abstraction and the operation of Naglu dam.

This information has been provided accordingly in the text.
"Similar to most other hydrological models (Efstratisdis et al., 2008), HYMOD_DS is not designed to model water abstractions for agricultural lands and dam operations within the basin. According to World Bank (2010), water demand for agricultural use is about 2,000 MCM (million cubic meters), or about 8.3% of the total annual flow. The Naglu dam (Figure 1) upstream of the Daronta streamflow gage forms the largest and most important reservoir in the basin, with an active storage of 379 MCM. In our hydrologic modelling process, the water consumed by irrigated croplands is implicitly accounted for by the evapotranspiration module. We note that the degree of irrigation impact during the time frame used for calibration (1960-1981) is likely much smaller than the current level. We also expect that using monthly data for calibration somewhat reduces the bias from human interference, particularly the daily operations of Naglu dam. Nevertheless, the calibration results for the gage below this dam (Daronta), and to a lesser extent the basin outlet (Dakah), should be approached with caution. Given that a majority of the gages examined in this study are on an underdeveloped branch of the Kabul River, issues of human interference on calibration are somewhat mitigated."

- P10278 L21 Should it not be "a genetic algorithm" as there are many kinds of genetic algorithms available for model calibration? Or you should state "the genetic algorithm introduced by Wang et al. 1991".

We made this clearer as suggested.

- P10279 L5 I wonder how these monthly streamflow values were calculated if not from daily measurements. If there are only monthly data available, I also wonder if the NSE is the best choice for goodness of fit criteria. Nevertheless, I like the argumentation given for choosing NSE but suggest to also mentioning here the use of KGE as another goodness of fit criterion

for model evaluation (so far, KGE is introduced in chapter 5 in the discussion and not in the methods section).

Unfortunately, the only observations that are available for public use are monthly. There is a report (Olson and Williams-Sether, 2010) clarifying that each monthly streamflow is the mean of the daily values for the month, and monthly values are calculated from daily values for all complete months of record. However, the daily values are not made available because there are political issues surrounding the trans-boundary use of the river's waters and potential projects planned on the river.

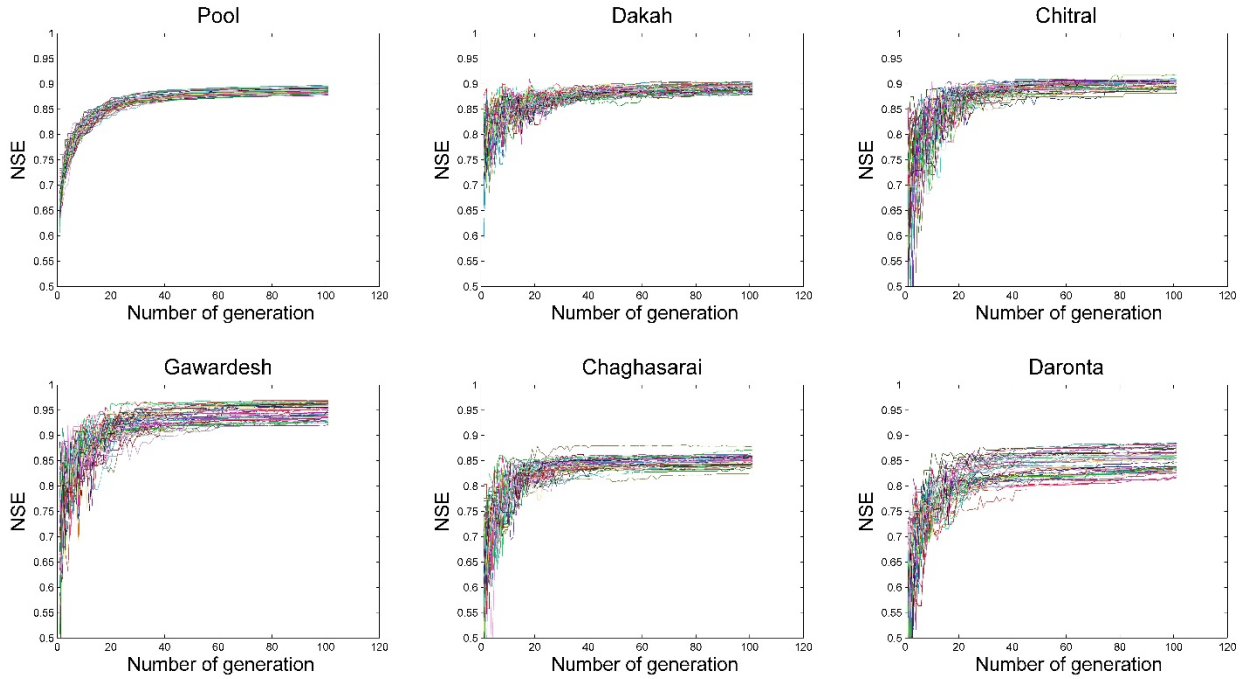
We have added the following details in the manuscript to clarify the immediate question regarding the data:
"Streamflow data were not collected in Afghanistan after September 1980 until recently because stream gaging was discontinued soon after the Soviet invasion of Afghanistan in 1979 (Olson and Williams-Sether, 2010). Though measurements were taken at a daily time step, data are only made available for public use at monthly aggregated levels, calculated using the mean of the daily values."

We acknowledged the limitation of the use of NSE for a model evaluation metric by writing this:
"However, in this particular study daily hydrologic model simulations can only be compared against available monthly streamflow records, reducing the number of viable objectives against which to calibration. That is, statistics representing peak flows, extreme low flows, and other daily flow regime characteristics often used in multi-objective optimization approaches are unavailable. We believe that the use of a monthly NSE value as a single objective, while coarse, does not inhibit our ability to provide insight into the research questions posed."

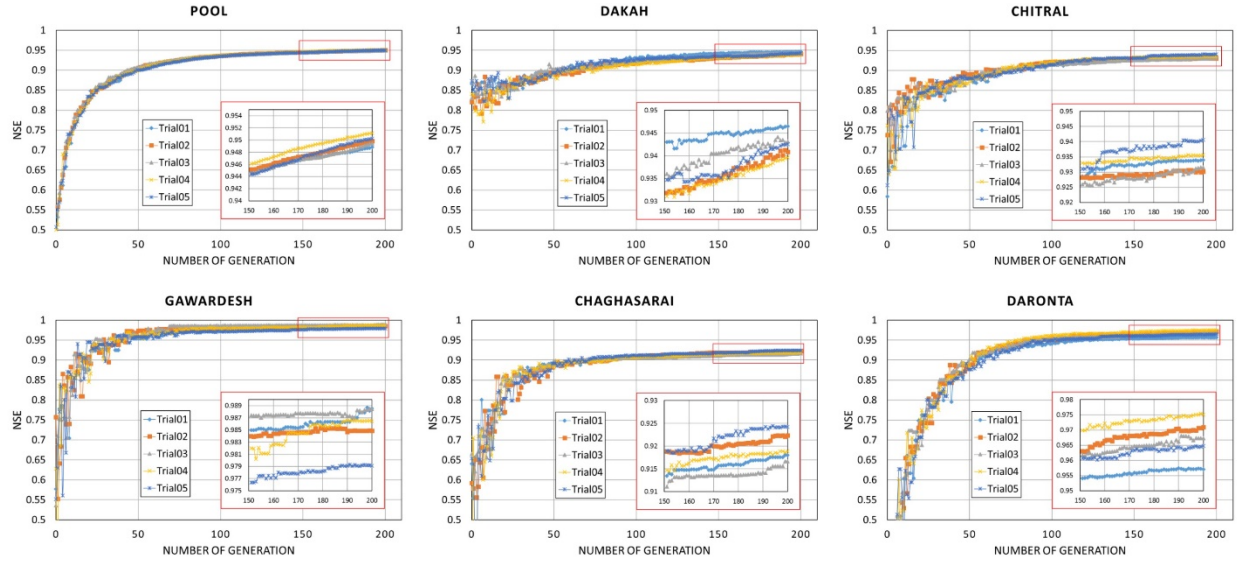
Also, we now introduce the KGE earlier in the Methods section to make clear that we are considering more than just the NSE for model diagnostics.

- P10282 L3 Are the numbers correct? The page before you present 15, 75 and 2400 parameter values being searched for in the various spatial set ups. Should it then not be 15x100 and 75x100? And why is 2400 multiplied by 200 and not by 100 as the others? Even though you state in the next sentence that the population/generation sizes were supported by convergence tests, the generation of numbers given here remains unclear.

We set up different numbers of population and generation in the GA algorithm according to the complexity of parameterization scheme. For instance, for the lumped parameterization, the number of parameter to be optimized is 15 and we considered 150 parameter sets. Those 150 parameter sets evolve through 100 generations, and the result of our convergence test showed a convergence while going through 100 generations. For the distributed parameterization scheme, there are more number of parameters to be calibrated. We considered 2400 parameter sets to calibrated 2400 parameters. Although it can be argued that having 2400 parameter sets to optimize 2400 parameters is not enough, we confirmed from the convergence test that this calibration setup shows a convergence behavior with 200 generations. Below, we enclosed the convergence test results.



GA convergence for the semi-distributed parameterization scheme with 750 parameter sets (population) and 100 iteration (generation)



GA convergence for the distributed parameterization scheme with 2400 parameter sets (population) and 200 iteration (generation)

- P10283 L11 step-wise (not step-wide)

Done.

- P10284 L12 the period “1960-1981” better covers all available discharge measurements given in Table 1.

  Yes, you are right. We changed it.

- P10294 L6 is shown in: : : (not was shown)

  We corrected it.

- Section 6 Conclusion P10295 L8 until P10296 L16 This is an extended summary of the results presented rather than a conclusion of the work. I think more effort should be put into real conclusions – what do we learn from the study, what are suggestions for future research, are results transferable to other regions or modelling approaches?

  As suggested, we tried to focus on the points that should be addressed in the conclusion part.

- Sections 5.2 and 5.3 The model performances for the upper subcatchments Kama and Asmar are generally very good. This is the same for Dakha (Figs 6 and 7). Glaciers have the largest extend in these subcatchments and I assume that they therefore contribute large volumes of water to total discharge at Dakah. Further, I assume that western catchments contribute only minor to total discharge as rainfall input is comparatively low (information on specific discharges for the various subcatchments would be helpful for a quick comparison). As you optimize your model using NSE, with NSE putting emphasis in matching peak flows, it does not come as a surprise to obtain good results for Dakah as long as subcatchments Kama and Asmar are calibrated sufficiently well.

  We updated Table 1 with more contents including the information on specific discharges for the sub-watersheds.
  In our study, we always treated Kama and Asmar as ungagged sub-watersheds, which means that we never tried to calibrate those two sites. All the available data at those sites were used for the validation purpose only. Dakah (the basin outlet) is the one against which the model calibrated. One of the main ideas we try to show in Sections 5.2 and 5.3 is that the calibration based on only the basin outlet does not provide a good performance at Kama and Asmar, while the pooled calibration does.

- Furthermore, the model performance of the ungauged sites Kama and Asmar are often very similar. Looking at the choice of stations that you treated ungauged and the general location of available gauging stations, I wonder why you have selected the Kama and Asmar, which belong to the same eastern area of the catchment. Why have you not selected the one in the west as a second interior test station (i.e. Daronta), or at least two subcatchments which are not draining into each other (e.g. Chaghasari and Asmar) and therefore being more independent than Kama and Asmar.

  The Government of Afghanistan with the support of the international donors (e.g. The World Bank) has developed comprehensive plans for the development of new hydro-power projects,

irrigation schemes and rehabilitation of old schemes on various rivers including the Kabul River (IUCN, 2010). Recently, Afghanistan and Pakistan reached an agreement in working on a 1,500MW hydropower project on Kunar River as part of the joint management of common rivers between the two countries (DAWN, 2013). For this study, Kama and Asmar were chosen and treated as ungagged sites in the processes of multisite calibrations because they align with the potential dam project.

This information has been provided accordingly in the text.
"The Government of Afghanistan has developed comprehensive plans for new hydropower projects on the Kabul River owing to its advantageous topography for the development of water storage and hydropower (IUCN, 2010), and recently reached an agreement with the Pakistan government to work on a 1,500MW hydropower project on the Kunar River (one of major tributary in the Kabul River basin) as part of the joint management of common rivers between the two countries (DAWN, 2013). …
Kama and Asmar stations are treated as ungaged sites because they align with the potential dam project on the Kunar River tributary."

- Section 5.4 Do you assume constant glacier volume to be discharging or are glaciers prone to glacier melt, resulting in smaller volume and spatial extend in the future and during your climate change simulation period. What are the expectations in glacier extend for the end of your simulation period in your catchment? Are calibrated model parameters still valid under these new boundary conditions? I expect not, as glacier melt is an important process, described by various parameters (Table 2) and needs rigorous calibration.

The hydrologic model (HYMOD_DS) used in this study does account for the changes in volume but has no ability to trace explicitly the spatial extend of glaciers. At the beginning of the simulation we were informed by the glacier volume (the amount water stored in the glaciers) which is provided by RGI3.2 and the area-volume relationship. A simple and possible way to trace the glacier extend from this study is to back-calculate the area with volume remaining at the end of simulation using the area-volume relationship. The model parameters related to the temperature-index glacier model stay the same once those are calibrated. Therefore, water from glacier melt with respect to a temperature above the threshold temperature will be same as long as glacier keep existing. We agree that it is hard to expect the calibrated parameters to be valid under new glacier conditions.
For our 20-year historical model simulation, we checked that the glacier volume decreases due to the ablation of glaciers larger than accumulation in the sub-watersheds that produce annual total flow larger than annual total precipitation as shown in the new Table 1. We argue that the high ratio of streamflow to precipitation is unrealistic and might be caused by error in precipitation data used in this study since precipitation measurement in high mountain areas is highly uncertain (Immerzeel et al., 2014). What we checked for the 20-year historical simulation and 30-year future simulation is that glaciers still stored enough water at the end of the simulations.
In our discussion for future work, we note the necessity of exploiting remote sensing and satellite products with which the evaluation of distributed hydrologic models with respect to model internal processes (e.g. snow, evapotranspiration, and glacier) becomes possible.

- S2 Please describe the meaning of abbreviations in the legend or figure caption

  We put the description in the figure caption.

- S8 Is this a simulation of the 100 yr flood event, at least this is what I understand from the text (P10294 L6 and following).

  We assumed that the reviewer meant Figure S6, not S8.
  No, this figure is showing the variability of optimum parameters derived from 50 trials of semi-distributed and distributed pooled calibrations. Here, we tried to explore the variability of 100-year flood estimates using 50 calibrated parameter sets for each calibration approach. Specifically, every time when the model was run with an optimum parameter set, we estimated the 100-year flood using the Log-Pearson III distribution for three locations (the basin outlet and 2 ungagged sites). With 50 100-year flood estimates for each calibration approach, we then examined the influence of the parameter variability on the flood estimates by comparing the flood estimates resulting from two calibration approaches.

Thank you.

References

While we were revising our manuscript, references listed below were added accordingly in the text.

Ahmad, M., and Wasiq, M.: Water resources development in Northern Afghanistan and its implications for Amu Darya Basin, The World Bank, Washington, D.C., 2004.

Boscarello, L., Ravazzani, G., and Mancini, M.: Catchment multisite discharge measurements for hydrological model calibration, Procedia Environmental Sciences, 19, 158-167, 2013.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the stregths of manual and automatic methods, Water Resources research, 36(12), 3663-3674, 2000.

Breuer, L., Huisman J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H. G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUChEM). I: Model intercomparison with current land use, Advances in Water Resources, 32, 129-146, 2009

Brown, C., Ghile, Y., Laverty, M., and Li, K.: Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector, Water Resources Research, 48, W09537, 2012.

Chen, J., Brissette, F. P., Poulin, A., and Leconte, R.: Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resources Research, 47, W12509, 2011.

DAWN: Pakistan, Afghanistan mull over power project on Kunar River, available at: http://www.dawn.com/news/1038435, last access: 2 January 2015, 2013.

Efstratiadis, A., Nalbantis, I., Koukouvinos, A., Rozos, E., and Koutsoviannis, D.: HYDROGEIOS: a semi-distributed GIS-based hydrological model for modified river basins, Hydrol. Earth Syst. Sci., 12, 989-1006, doi:10.5194/hess-12-989-2008, 2008.

Exbrayat, J. F., Buytaert, W., Timbe, E., Windhorst, D., and Breuer, L.: Addressing sources of uncertainty in runoff projections for a data scarce catchment in the Ecuadorian Andes, Climatic Change, 125, 221-235, 2014.

FAO: Global Land Cover Share Database version 1.0, available at: http://www.fao.org/geonetwork, last access: 2 January 2015, 2013.

Federer C. A., Vorosmarty C., Fekete B.: Intercomparison of methods for calculating potential evaporation in regional and global water balance models. Water Resour Res 32:2315–2321, 1996.

Flugel, W. A.: Delineating Hydrological Response Units (HRU's) by GIS analysis for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Brol, Germany, Hydrol. Processes, 9, 423-436, 1995.

Forsythe, W. C., Rykiel Jr., E. J., Stahl, R. S., Wu, H., Schoolfield, R. M.: A model comparison for daylength as a function of latitude and day of year, Ecological Modelling, 80, 87-95, 1995.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Towards improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751-763, 1998.

Immerzeel, W. W., Petersen, L., Ragettli, S., and Pellicciotti, F.: The importance of observed gradients of air temperature and precipitation for modeling runoff from a glacierized watershed in the Nepalese Himalayas, Water Resour. Res., 50, 2212–2226, 2014.

IUCN: Towards Kabul Water Treaty: Managing Shared Water Resources – Policy Issues and Options, IUCN Pakistan, Karachi, 11 pp, 2010.

Koutsoyiannis, D., Efstratiadis, A., Mamassis, N., and Christofides, A.: On the credibility of climate predictions, Hydrological Sciences Jounal, 53(4), 671-684, 2008.

Lerat, J., Andreassian V., Perrin, C., Vaze, J., Perraud J. M., Ribstein, P., and Loumagne C.: Do internal flow measurements improve the calibration of rainfall-runoff models?, WATER RESOUR RES, 48, W02511, 2012.

Lu J, Sun G, McNulty S. G., Amataya D. M.: A comparison of six potential evapotranspiration methods for regional use in the southeastern United States. J Am Water Resour Assoc 3:621–633, 2005.

Olson, S. A., and Williams-Sether, T.: Streamflow characteristics at streamgages in Northern Afghanistan and selected locations, U. S. Geological Survey, Reston, Virginia, 2010.

Perez, J., Menendez, M., Mendez, F. J., and Losada, I. J.: Evaluating the performance of CMIP3 and CMIP5 global climate models over the north-east Atlantic region, Climate Dynamics, 43, 2663-2680, 2014.

Shakir, A. S., Rehman, H., and Ehsan, S.: Climate change impact on river flows in Chitral watershed, Pakistan Journal of Engineering and Applied Sciences, 7, 12-23, 2010.

Steinschneider, S., Wi, S., and Brown, C.: The integrated effects of climate and hydrologic uncertainty on future flood risk assessments, Hydrological Processes, DOI: 10.1002/hyp.10409, 2014.

USDA-NRCS: Global Soil Regions Map and Global Soil Suborder Map Data from US Department of Agriculture, Natural Resource Conservation Service, 2007.

Vorosmarty C. J., Federer C. A., Schloss A. L.: Potential evaporation functions compared on US watersheds: possible implications for global-scale water balance and terrestrial ecosystem modeling. J Hydrol 207:147–169, 1998.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrology and Earth System Sciences, 5(1), 13-26, 2001.

World Bank: Afghanistan – Scoping strategic options for development of the Kabul River Basin: a multisectoral decision support system approach, World Bank, Washington, D. C., 2010.